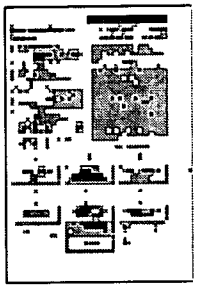
Title: **JP2002269139A2: METHOD FOR RETRIEVING DOCUMENT**

Derwent Title: Document search method involves searching document based on divided character sequence index and word index and designating document, when character sequence equals corresponding word [Derwent Record]



Country: **JP** Japan

Kind: **A2** Document Laid open to Public inspection

Inventor: **OGAWA YASUTSUGU**;

Assignee: **RICOH CO LTD**
News, Profiles, Stocks and More about this company

Published / Filed: **2002-09-20** / 2001-03-08

Application Number: **JP2001000064404**

IPC Code: **G06F 17/30;**

Priority Number: 2001-03-08 **JP2001000064404**

Abstract: PROBLEM TO BE SOLVED: To easily and fast retrieve a document including a designated character string from a registered document group.

SOLUTION: This document retrieving method comprises a text dividing means for disassembling text being a registered document or a retrieval character string into n-grams (n character set) and words, an n-gram index for holding appearance information about n-grams in the registered document in each n- gram, a word boundary index for holding appearance information about a word boundary in the registered document, a character string unit retrieving means for retrieving a document including the retrieval character string or an appearance position in the document by referring to the n-gram index on the basis of results obtained by dividing the retrieval character string to the n-grams, and a word unit retrieving means for deciding whether the retrieval character string appears as a word by referring to the word boundary index on the basis of results obtained by dividing the retrieval character string into words with respect to results of the character string unit retrieving means and retrieving a document including the retrieval character string as a word.

Family: None

Other Abstract Info: DERABS G2003-005458

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－269139
（P2002－269139A）
(43)公開日　平成14年９月20日(2002.9.20)

(51)Int.Cl.⁷	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/30	３３０	Ｃ０６Ｆ　17/30	３３０Ｃ　５Ｂ０７５
	１７０		１７０Ａ

審査請求　未請求　請求項の数７　ＯＬ　（全　６　頁）

(21)出願番号　特願2001－64404(P2001－64404)

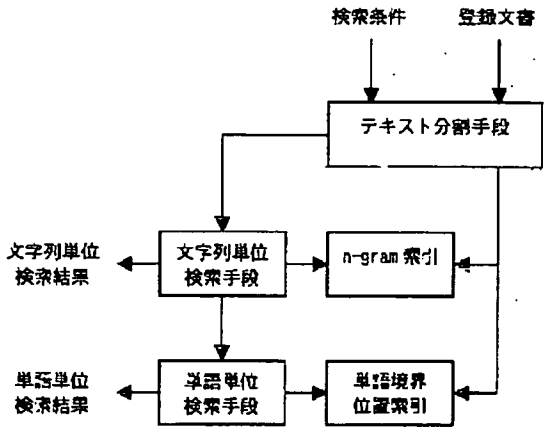
(22)出願日　平成13年３月８日(2001.3.8)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　小川　泰嗣
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(74)代理人　100101177
弁理士　柏木　慎史　（外２名）

Ｆターム(参考)　5B075 ND03 PR04 QM02

(54)【発明の名称】　文書検索方法

(57)【要約】
【課題】　登録された文書群から指定された文字列を含む文書を簡易にかつ高速に検索することである。
【解決手段】　登録文書あるいは検索文字列であるテキストをｎ－ｇｒａｍ（ｎ文字組）および単語に分解するテキスト分割手段、登録文書中のｎ－ｇｒａｍに関する出現情報をｎ－ｇｒａｍごとに保持しておくｎ－ｇｒａｍ索引、登録文書中の単語境界に関する出現情報を保持しておく単語境界索引、検索文字列をｎ－ｇｒａｍに分割した結果に基づいてｎ－ｇｒａｍ索引を参照して検索文字列を含む文書あるいはその文書における出現位置を検索する文字列単位検索手段、文字列単位検索手段の結果に対し検索文字列を単語に分割した結果に基づいて単語境界索引を参照して検索文字列が単語として現われているかを判断し、検索文字列を単語として含む文書を検索する単語単位検索手段、からなる。



【特許請求の範囲】
【請求項１】　登録文書あるいは検索文字列であるテキストをｎ－ｇｒａｍ（ｎ文字組）および単語に分解するテキスト分割手段、登録文書中のｎ－ｇｒａｍに関する出現情報をｎ－ｇｒａｍごとに保持しておくｎ－ｇｒａｍ索引、登録文書中の単語境界に関する出現情報を保持しておく単語境界索引、検索文字列をｎ－ｇｒａｍに分割した結果に基づいてｎ－ｇｒａｍ索引を参照して検索文字列を含む文書あるいはその文書における出現位置を検索する文字列単位検索手段、文字列単位検索手段の結果に対し検索文字列を単語に分割した結果に基づいて単語境界索引を参照して検索文字列が単語として現われているかを判断し、検索文字列を単語として含む文書を検索する単語単位検索手段、からなることを特徴とする文書検索方法。
【請求項２】　ｎ－ｇｒａｍ索引および単語境界索引は、出現情報として出現文書の文書識別子・出現文書での出現回数・出現文書での出現位置を含むことを特徴とする請求項１記載の文書検索方法。
【請求項３】　単語単位検索手段は、検索文字列の文書中での出現の先頭位置と末尾位置がその文書の単語境界に含まれている文書を、検索文字列を単語として含む文書として検索することを特徴とする請求項１又は請求項２記載の文書検索方法。
【請求項４】　単語単位検索手段は、検索文字列の文書中での出現の先頭位置がその文書の単語境界に含まれている文書を、検索文字列ではじまる単語を含む文書として検索することを特徴とする請求項１又は請求項２記載の文書検索方法。
【請求項５】　単語単位検索手段は、検索文字列の文書中での出現の末尾位置がその文書の単語境界に含まれている文書を、検索文字列で終わる単語を含む文書として検索することを特徴とする請求項１又は請求項２記載の文書検索方法。
【請求項６】　単語境界索引は、単語の長さごとに単語境界情報を記録することを特徴とする請求項１又は請求項２記載の文書検索方法。
【請求項７】　単語境界索引は、特定の品詞に属する単語に関する単語境界情報を記録することを特徴とする請求項１又は請求項２記載の文書検索方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、登録された文書群から指定された文字列を含む文書を検索する文書検索方法に関する。
【０００２】
【従来の技術】登録された文書群から必要な文書を検索する文書検索方法には、文字列単位の検索（以下、文字列単位検索）と単語単位の検索（以下、単語単位検索）の２つの方法がある。文字列単位検索では、ユーザが指定した文字列（以下、検索文字列）を文字列として含む文書を検索する。文字列単位検索を高速化する方法としては、文書中のｎ文字組（以下、ｎ－ｇｒａｍ）を索引単位として、索引単位ごとに出現した文書識別子と文書内での出現位置を記録する　ｎ－ｇｒａｍ　索引を用意しておく方法が知られている。一方、単語単位検索では、ユーザが指定した検索文字列を単語として含む文書を検索する。単語単位検索を高速化する方法としては、文書中の単語を索引単位として、索引単位ごとに出現した文書識別子と文書内での出現位置を記録する単語索引を用意しておく方法が知られている。
【０００３】
【発明が解決しようとする課題】しかし、いずれの検索方法にも問題がある。文字列単位検索の場合、単語境界を無視して検索するため、ユーザが意図しない文書が検索結果に含まれることがある。例えば、「帯電」（電気を帯びること；ｅｌｅｃｔｒｉｆｉｃａｔｉｏｎ）を検索文字列とした場合、「彼女の携帯電話」という文書が検索されてしまう。一方、単語単位検索の場合、日本語においては単語の区切れが明示的に示されないため、索引を作成する際に形態素解析などを施して単語を切り出す必要があるが、現在の技術レベルでは形態素解析の誤りが避けられない。したがって、形態素解析誤りが原因で検索漏れが起こる。例えば、「東京都にある清水寺」という文書を登録する際、／東／京都／に／ある／清水寺／と形態素解析されるべきところを／東京／都／に／ある／清水／寺／のように誤って解析されれば、検索文字列が「京都」のときに「東京都にある清水寺」という文書を検索することができない。
【０００４】上述した問題を避けるには、システムが両検索方法をサポートし、ユーザがニーズに応じていずれかの検索方法を指定できるようにすればよい。そのような検索方法の従来技術として特開２０００－６７０７０がある。この発明では、文書登録時に単語の切れ目に特殊な区切り文字を挿入し、区切り文字を挿入したデータからｎ－ｇｒａｍを切り出し、索引を作成する。その際、区切り文字の前後を連結させて得られるｎ－ｇｒａｍも切り出して索引に記録する。ユーザが単語単位検索を指定した場合には区切り文字を含むｎ－ｇｒａｍを無視することなく検索処理を行い、文字列単位検索を指定した場合には区切り文字を含むｎ－ｇｒａｍを無視して検索処理を行う。
【０００５】別の従来技術としては特開平７－８５０３３号公報に記載された技術がある。この発明では、文字ごとにその文字が出現する文書、その文書における出現位置、および各出現位置が単語の先頭／末尾かのフラグを記録しておく。検索時には文字ごとの出現位置に基づいて文字列単位の検索を実現するとともに、単語の先頭／末尾かのフラグも参照することで単語単位の検索も実現する。

【0006】しかし、前者には以下の問題がある。単語の切れ目を区切り文字で表現している。一般に文字は固定長（例えば、UCS2表現のユニコードでは2バイト）で表現されるが、表現可能な値がすべて意味ある文字として使われる場合には、この方法を適用できない。

【0007】一方、後者には以下の問題がある。文字列単位の検索が文字索引に基づいているので、n-gram索引と比較して検索速度が遅い。

【0008】さらに、両者に共通の問題点として以下の問題もある。単語の切れ目を発見するための形態素解析系（あるいはそれが使用する辞書）を更新した場合、切れ目の位置が変わるため、索引全体の作り直しが必要である。その結果、索引のメンテナンス作業に多大な時間を要する。

【0009】

【課題を解決するための手段】本発明はこれら問題点を鑑みて発明されたもので、文字列単位検索用のn-gram索引に加えて、単語の切れ目の位置を記録した単語境界索引を用意する。文字列単位検索は従来と同じくn-gram索引を用いて処理する。単語単位索引は、文字列単位検索を行った上で検索文字列の文書中での出現位置が単語境界に一致するかを単語境界索引を用いて判定し、実現する。本発明によれば、特殊な区切り文字が不要なので、どんな文字コードに対して適用可能である。また、文字列単位検索をn-gram索引を用いて行うので高速である。さらに、単語の切れ目に関する情報はn-gram索引には影響しないので、形態素解析系を更新した場合でも単語境界索引のみを作り直せばよい。

【0010】

【発明の実施の形態】本発明の第1の実施の形態を図面に基づいて説明する。図1において、テキスト分割手段は、登録文書あるいは検索条件に含まれるテキストをn-gramおよび単語に分解する。N-gram索引は、登録文書を分割したn-gramの情報を、検索のために保持するものである。単語境界索引は、登録文書を分割した単語の情報を、検索のために保持するものである。文字列単位検索手段は、テキスト分割手段が検索文字列を分割したn-gramに基づいて、n-gram索引を用いて検索文字列を含む文書を検索するものである。単語単位検索手段は、文字列単位検索手段で求められた文字列検索結果において検索文字列が単語として出現しているかを単語境界索引を用いて決定するものである。

【0011】登録においては、文書が与えられるとテキスト分割手段でn-gramと単語に分割し、それら出現情報をn-gram索引および単語境界索引に記録する。

【0012】なお、単語への分割には形態素解析を利用するが、形態素解析は既存の手法（例えば、松本裕治他、「単語と辞書」言語の科学第3巻、岩波書店の53ページ以降に書かれている）を採用すればよい。

【0013】登録手順を図2の例を用いて説明する。登録文書（＝文書1）の内容が図2の（a）、その形態素解析結果は（b）の通りであるとする。いま、索引単位をbi-gram（2文字組；n＝2のn-gram）とすると、この文書は「携帯」「帯電」のようなbi-gramに分割され、（c）のようなn-gram索引ができる。ここで、左側の文字列（「携帯」など）が索引単位であるbi-gramを表し、右側の数字がその索引単位が出現した文書識別子、その文書での出現回数、各出現位置（文書先頭を1とした場合の文字数）を表す。例えば、「帯電」に対する{1，1，（5）}は、文書1には1回出現し、その位置は5文字目であることを意味する。（d）は単語境界索引で、形態素解析結果で得られる単語境界の出現位置を記録したものである。データの記述方法はn-gram索引と同じであり、{1，5，（1，3，4，6，8）}は文書1には5回出現し、その位置は1，3，4，6，8文字目であることを意味する。最後の8文字目は最後の単語の末尾位置に対応する。

【0014】なお、n-gram索引の構成方法は、n＝2であるbi-gram 以外を用いてもかまわない。さらには、文字種に応じてnを調整する方法でもかまわない。また、索引中では文書識別子などを圧縮〔例えば、I. H. Witten他、Managing Gigabytes （second edition），Morgan Kaufmannの114～128ページの方法〕して記録してもよい。

【0015】文字列単位検索では、検索文字列が与えられると、テキスト分割手段はn-gramに分割し、文字列単位検索手段は分割されたn-gramに関する登録文書中の出現文書あるいは出現文書とその文書内での出現位置を用いて検索文字列を含む文書を決定する。

【0016】図2の索引を用いるとして、文字列単位検索処理を説明する。検索文字列を「帯電」とすると、（この検索文字列自体がbi-gramなので）テキスト分割手段は「帯電」を抽出する。つぎに、文字列単位検索手段は、索引を調べると、「帯電」は文書1に出現していることがわかり、文書1を検索結果とする。検索文字列が「携帯電話」であれば、テキスト分割手段は「携帯」［帯電］「電話」の3つのbi-gramを抽出する。つぎに、文字列単位検索手段は、これらのbi-gramを全て含む文書を特定し、その文書においてbi-gramが連続した位置に出現している場合にはその文書を検索結果とする。この場合、「携帯」［帯電］［電話］の出現位置は4，5，6と1ずれているので「携帯電話」は文書1において出現位置4に現われていると判断でき、文書1を検索結果とする一方、単語単位検索では、文字列単位検索において文字列検索手段が

求める検索文字列の文書における出現が単語としてであるかを判定する。手順は以下の通りである。

【００１７】（１）検索文字列を形態素解析し、単語の区切りを得る。

【００１８】（２）検索文字列で文字列検索を行い、検索文字列を含む文書を特定する（なお、（３）（４）から戻ってきた場合には、検索文字列を含む次の文書を特定する）。見つからなければ終了。

【００１９】（３）前述の（２）で検索された文書について、検索文字列の出現位置を得る（なお、（４）から戻ってきた場合には、検索文字列の次の出現位置を得る）。見つからなければ（２）に戻る。

【００２０】（４）前述の（３）で得られた出現位置の先頭から末尾までに含まれる単語境界を単語境界索引から得る。その相対位置が（１）で得られた検索文字列の単語境界と一致すれば（２）で特定された文書を検索結果に追加し、（２）に戻る。相対位置が検索文字列の単語境界と一致しなければ、（３）に戻る。

【００２１】検索文字列「帯電」を例に説明する。まず、形態素解析し／帯電／という結果が得られ、単語境界は１，３文字目（先頭位置が１文字目で、末尾位置は先頭位置に単語の長さ２を足して得られる）とわかる。次に文字列検索すると文書１が特定され、そこでの出現位置は５文字目から７文字目とわかる。ところが、この文書における単語境界は…，４，６，…文字目で、一致しないことがわかる。これ以外の出現位置・文書は見つけられないので、単語単位検索によれば該当文書なしという検索結果になる。つまり、「帯電」は単語としては現われていないことがわかる。

【００２２】検索文字列が「携帯電話」だと以下のようになる。まず、形態素解析し／携帯／電話／という結果が得られ、単語境界は１，３，５文字目とわかる。次に文字列検索すると文書１が特定され、そこでの出現位置は４文字目から８文字目とわかる。一方、この出現位置付近の単語境界は４，６，８文字目であり、検索文字列における単語境界と一致する。したがって、文書１は検索結果に含まれる。この方法では、単語境界を文書の先頭からの文字数で表現しているので、特殊文字を使用する必要がなく、任意の文字コードに対して適用可能である。また、単語境界は文字列検索用のｎ－ｇｒａｍ索引とは別に作成・管理されるので、形態素解析系の変更時には単語境界索引だけを作り直せばよく、索引のメンテナンス作業が軽減される。

【００２３】本発明の第２の実施の形態を説明する。前述の第１の実施形態では、検索文字列の単語区切りと文字列単位検索結果で得られる文書中の出現位置範囲の単語区切りが一致することを検査する。したがって、検索文字列が長くて単語区切りが多く含まれる場合には検索文字列と文書中での単語区切りの一致検査に要する処理量も多くなり、検索に時間を要する。

【００２４】そこで、検索文字列と文書中での単語境界の一致検査で全ての単語境界を用いるのではなく、先頭位置と末尾位置のみを使用する。検索文字列が３個以上の単語境界を含むのは複合語と考えられるが、ほとんどの場合、先頭位置と末尾位置の単語境界が一致すれば文字列単位検索で生じた誤検索を除くことができる。また、先頭位置と末尾位置しか一致を調べないのであれば検索文字列を形態素解析する必要もなく、一致検査も検索文字列の長さに依存しないので、検索を高速化できる。

【００２５】この方法では、単語単位検索の手順は以下のように置き換わる。

【００２６】（１）検索文字列で文字列検索を行い、検索文字列を含む文書を特定する（なお、（２）（３）から戻ってきた場合には、検索文字列を含む次の文書を特定する）。見つからなければ終了。

【００２７】（２）前述の（１）で検索された文書について、検索文字列の出現位置を得る（なお、（３）から戻ってきた場合には、検索文字列の次の出現位置を得る）。見つからなければ（１）に戻る。

【００２８】（３）前述の（２）で得られた出現位置の先頭から末尾が、単語境界索引に記録されていれば（１）で特定された文書を検索結果に追加し、（１）に戻る。記録されていなければ、（２）に戻る。

【００２９】検索文字列が「携帯電話」の場合、文字列検索される文書１における出現位置は４文字目から８文字目である。これらは単語境界索引に記録されているので、検索文字列の前後の位置は単語境界であり、文書１は検索結果に含まれる。

【００３０】本発明の第３の実施の形態を説明する。前述の第１、第２の実施の形態では単語単位検索により、検索文字列が単語として出現しているかを判断した上で検索を行っていた。しかし、特定の文字列で始まる単語を含む文書を検索したい場合もある（以下では前方一致検索と呼ぶ）。

【００３１】前方一致検索では、文字列単位検索において文字列検索手段が求める検索文字列の文書における出現の先頭が単語境界であるかを判定する。検索手順は以下の通りである。

【００３２】（１）検索文字列で文字列検索を行い、検索文字列を含む文書を特定する（なお、（２）（３）から戻ってきた場合には、検索文字列を含む次の文書を特定する）。見つからなければ終了。

【００３３】（２）前述の（１）で検索された文書について、検索文字列の出現位置を得る（なお、（３）から戻ってきた場合には、検索文字列の次の出現位置を得る）。見つからなければ（１）に戻る。

【００３４】（３）前述の（２）で得られた出現位置の先頭が、単語境界索引に記録されていれば（１）で特定された文書を検索結果に追加し、（１）に戻る。記録さ

れていなければ、（２）に戻る。

【００３５】「携帯」ではじまる単語を含む文書を特定したいという場合を例に説明する。「携帯」で文字列単位検索される文書１における出現位置の先頭は４文字目である。これは単語境界索引に記録されているので、文書１は検索結果に含まれる。

【００３６】検索文字列が「帯電」であれば、その文字列単位検索で得られる出現位置の先頭は５文字目で、単語境界索引に記録されていないので文書１は検索結果に含まれない。

【００３７】本発明の第４の実施の形態を説明する。第３の実施の形態とは異なり、特定の文字列で終わる単語を含む文書を検索したい場合もある（以下では、後方一致検索と呼ぶ）。

【００３８】後方一致検索では、文字列単位検索において文字列検索手段が求める検索文字列の文書における出現の末尾が単語境界であるかを判定する。検索手順は以下の通りである。

【００３９】（１）検索文字列で文字列検索を行い、検索文字列を含む文書を特定する（なお、（２）（３）から戻ってきた場合には、検索文字列を含む次の文書を特定する）。見つからなければ終了。

【００４０】（２）前述の（１）で検索された文書について、検索文字列の出現位置を得る（なお、（３）から戻ってきた場合には、検索文字列の次の出現位置を得る）。見つからなければ（１）に戻る。

【００４１】（３）前述の（２）で得られた出現位置の末尾が、単語境界索引に記録されていれば（１）で特定された文書を検索結果に追加し、（１）に戻る。記録されていなければ、（２）に戻る。

【００４２】「電話」で終わる単語を含む文書を特定したいという場合を例に説明する。「電話」で文字列単位検索される文書１における出現位置の末尾は８文字目である。これは単語境界索引に記録されているので、文書１は検索結果に含まれる。

【００４３】本発明の第５の実施の形態を説明する。第１の実施の形態では、全ての単語境界をまとめて記録していたため、検索時に単語境界索引から読み出し、照合処理にまわされるデータ量が多いという問題がある。通常の日本語であれば単語の平均文字数は３文字程度であるので、１０００文字の文書であれば単語数は３００程度となる。

【００４４】そこで、本実施の形態では、単語の長さに注目し、単語の長さ（文字数）によって単語境界位置を分類し、異なるレコードとして記録する。例えば、図２（ｄ）には単語長が１と２のものが含まれているので、単語長ごとに異なるレコードとした場合、図３のようになる。なお、図２（ｄ）では単語境界位置の末尾に最後の単語の末尾位置を記録していたが、この方法では単語の先頭位置と長さから単語の末尾位置が求められるので、最後の単語の末尾位置を記録する必要はない。したがって、単語長１と２の単語境界の出現回数の合計は４　であり、図２（ｄ）の場合の出現回数５よりも１少なくなっている。

【００４５】この方法で記録した場合でも、単語単位検索の流れは同じである。ただし、ステップ（４）で文書から単語境界位置を読み出す際には、検索文字列中の単語の長さに応じた境界位置を読み出す点が異なる。例えば、検索文字列が「帯電」、「携帯電話」であれば、いずれも検索文字列中の単語の長さは２なので、単語長２に対応する単語境界位置データのみを使用する。

【００４６】一方、第２、第３、第４の実施の形態に示した単語単位検索では照合すべき単語境界に対応する単語の長さが一意に特定できないので、全ての長さの単語境界位置を位置順にマージした結果を単語境界位置として使用する必要がある。マージの際、最後の単語の末尾位置を最後の単語の先頭位置と長さから計算し、マージ結果に含める必要がある。図３のデータであれば、（３）と（１，４，６）を単にマージして（１，３，４，６）とするのでなく、最後の単語の先頭位置６にその単語の長さ２を足した８をアペンドした（１，３，４，６，８）を単語境界の照合に使用する。

【００４７】本発明の第６の実施の形態を説明する。第１の実施の形態では、全ての単語境界を記録していたため、単語境界索引が大きいという問題がある。１文書中の単語数が３００の場合、位置を４バイトで記録すると、１文書あたり１．２キロバイト必要になる。出現位置は圧縮することでデータ量を削減することは可能だが、それでも記録すべき単語境界位置を減らすことが望ましい。

【００４８】そこで、本実施の形態では、検索語として実際に使用されるのは名詞等の自立語が大半であるという点に注目し、文書登録時に形態素解析結果から特定の品詞の単語についてのみ単語境界位置を単語境界索引に記録する。例えば、図４（ａ）の文書は形態素解析によって（ｂ）のように単語分割される。これを実施形態１の単語境界索引に記録すると（ｃ）のようになる。この文書の大半の単語は助詞・助動詞・形式名詞であり、検索文字列として使用されることが多い名詞類は「経験」「台風」だけである。それにもかかわらず（ｃ）では全ての単語の位置を記録しているので、多くの領域を使用している。これに対し、本実施形態では名詞類である「経験」「台風」の位置だけを記録する。この方法では記録される単語が連続しているとは限らないので、連続していない部分には連続していないことを表す特別な値として０を挿入している。

【００４９】なお、この例では選択する品詞を名詞としたが、それ以外の品詞を含めてもかまわない。

【００５０】

【発明の効果】請求項１および請求項２に記載された文

書検索方法においては、文字列単位検索用のｎ－ｇｒａｍ索引とは別個の単語境界索引を用いて単語単位検索を提供しているので、どんな文字コードに対しても適用可能であり、検索が高速で、索引のメンテナンスが簡単である。

【００５１】請求項３記載の文書検索方法においては、単語単位検索時に照合すべき単語境界の個数が少なくなるので、検索処理が高速になる。

【００５２】請求項４記載の文書検索方法においては、検索文字列ではじまる単語を検索できるので、ユーザに柔軟な検索機能を提供できる。

【００５３】請求項５記載の文書検索方法においては、検索文字列でおわる単語を検索できるので、ユーザに柔軟な検索機能を提供できる。

【００５４】請求項６記載の文書検索方法においては、単語境界索引を単語長に応じて分割するので、単語単位検索時に参照すべき単語境界データが削減され、検索処理が高速になる。

【００５５】請求項７記載の文書検索方法においては、単語境界索引に記録する単語を品詞によって選択するので、単語境界索引が小型化される。

【図面の簡単な説明】

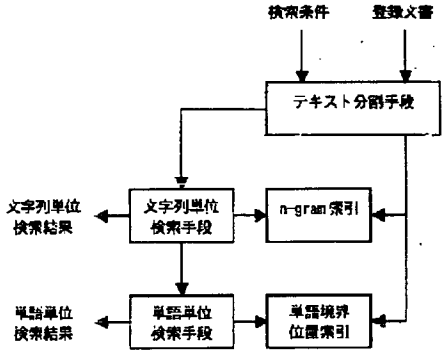
【図１】文書検索方法の概要を示すブロック図である。

【図２】登録文書と索引の例を示す説明図である。

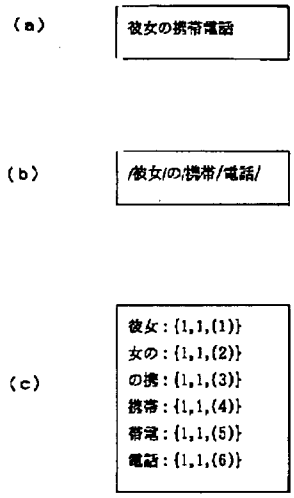
【図３】単語の長さで分割した単語境界索引の例を示す説明図である。

【図４】登録文書と単語境界索引の例を示す説明図である。

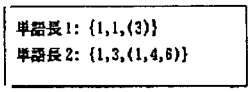
【図１】



【図２】



【図３】



【図４】